Two persons dialogue corpus made by multiple crowd-workers

Tetsuro Takahashi and Hikaru Yokono

Abstract Crowdsourcing is a promising method for corpus development. While crowdsourcing works well for simple independent tasks, it is not clear how to apply crowdsourcing to develop a corpus of dialogue which originally consists of continuous utterances which depend on the previous ones. This paper proposes a methodology to create a dialogue corpus assigning multiple crowd workers to one of two participants in a dialogue. The analysis based on the comparison with a corpus made in straightforward method showed that the quality of the corpus is comparable though the length of utterance and dialogue tends to be different.

1 Introduction

Corpora play an important role in research and development on Natural Language Processing including dialogue. They allow us to analyze linguistic phenomena and give us statistical information which enables to apply machine learning algorithms.

In this paper, we aim to develop a corpus of dialogue in which two participants have a text conversation. A straightforward method to develop such the corpus is to prepare two participants and let them to have a conversation. They don't have to be in the same place in a case a virtual environment such as a chat channel is provided. Even the participants talk on the virtual environment, we still have to pay the cost for arranging the pair and have the participants to work at the same time.

Crowdsourcing has been used to develop many kinds of data in recent years and has shown the capability in several aspects such as cost, latency and diversity. In the use of crowdsourcing there are still several options to develop a dialogue corpus as shown in Table 1. We propose a methodology for developing a dialogue corpus by the 3rd method in Table 1 which has cost efficiency. The 3rd one was 15 times

Tetsuro Takahashi, Hikaru Yokono
Fujitsu Laboratories Ltd., 4-1-1, Kamikodanaka Nakahara-ku Kawasaki Japan, e-mail: {takahashi.tet,yokono.hikaru}@jp.fujitsu.com

cheaper than the 1st one in our case. The 2nd one is not suitable for crowdsourcing because workers have to wait for the other one. This is for only the comparison.

In this paper, a “dialogue” means a sequence of utterances made by two participants, a “participant” means a role of a person in a dialogue, and a “worker” means a person who actually generate utterances.

Table 1 Three options to apply crowdsourcing to development of dialogue corpus

1.	Hire two workers for a dialogue, then let them to have a text chat in real time.
2.	Hire two workers for a dialogue, then let them to have a text chat in asynchronous timing in which a worker is assigned after the other worker made an utterance.
3.	Assign multiple workers to both sides of participant.

2 Related Work

One of the problems in dialogue corpus development is the cost for management of workers such as gathering and pairing them. Several studies address this issue: reducing the cost by some tools and utilizing a other data similar to real dialogue. Higashinaka et al. constructed a corpus consisting of chat dialogues between a human participant and the system with publicly available chat API [1]. Sugiyama et al. propose the method of utterance generation with Twitter data[7].

Crowdsourcing is widely used for some purposes in the research field of dialogue these days. Mitchell et al. developed a corpus of natural language generation templates by using crowdsourcing [3] . Paperno et al. constructed a data set evaluated by crowdsourcing for natural language understanding [6]. Lasecki et al. proposed a system that labels events in videos by crowd workers [2]. Regarding to their work as a corpus development, the corpus was built by real users and crowd workers. Our corpus was built by two participant roles played by multiple workers in both sides. It is important for crowdsourcing to maintain workers and data quality as several studies address this issue(cf. [5], [4]).

The contributions of this paper are i) to propose a methodology to create a dialogue corpus assigning multiple crowd workers to one of two participants in a dialogue and ii) to investigate the generated corpus comparing to a dialogue corpus developed in the straightforward method.

3 Pseudo dialogue corpus

We propose to decompose the task of dialogue construction into a set of individual tasks of utterance generation, and to assign multiple workers to both sides of participant in a dialogue. Since dialogue is originally a continuous work by participants, the proposed method needs a device to maintain the quality. For the purpose we showed workers i) The side of participant's role, ii) Background, and iii) Utterance sequence history. Given the information, the task is to generate an utterance as the specified role of participant based on the background and the utterance sequence history.

The topic of the dialogue corpus is "dialogues between a real-estate agent and a customer". We made the corpus in Japanese and all the instances in this paper are translated ones. Figure 1 shows the worker assignment to construct a dialogue. Each worker was assigned to only one side in a dialogue in order to prevent to show a background to agent-side workers.

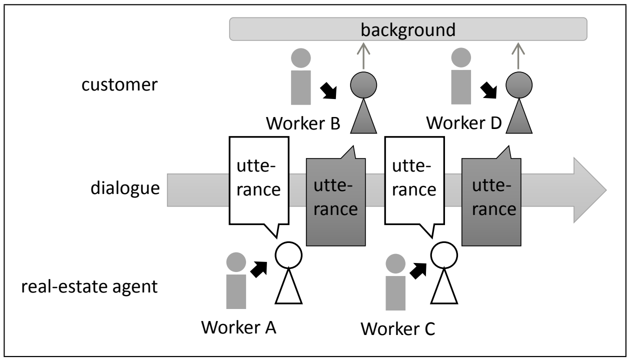

Fig. 1 Worker assignment for a dialogue construction

The backgrounds for customer-side workers were also collected via crowdsourcing for the reason of diversity. We asked workers who have experience of conversation in real-estate office and collected 100 background descriptions. We chose 10 out of the 100 descriptions. The following is the instance. We showed nothing for agent-side background.

> A customer is to live with her long-standing boyfriend and will move from her studio apartment. On this occasion, she wants to step up her cooking skill and wants to live a property which has a convenient kitchen with some ranges.

The goal of agent-side workers is to draw out customer's requests which are enough for searching houses. A dialogue closes in the following three cases.

- An agent-side worker judges that enough information has been derived from customer. This is successful completion.
- A worker judges that it is impossible to continue the dialogue anymore. "inadequate" tag is annotated by the worker.
- A number of utterances in the dialogue exceeds a limit (20 in the evaluation). "abort" tag is annotated automatically.

We call the dialogue made by the proposed method the **pseudo dialogue**. In order to evaluate the pseudo dialogue, we also constructed a dialogue corpus by crowdsourcing assigning one worker to one participant. Other settings are the same as described before. We call this dialogue the **chat dialogue**. This is correspond to the 2nd method in Table 1.

Table 2 Statistics of pseudo dialogue corpus

# of dialogue	270	average duration for dialogue (hour)	183.92
# of utterance in total	3,168	average interval for utterance (hour)	21.27
# of worker	965	# of dialogues tagged “inadequate”	88
# of pair of worker	2,696	# of dialogues tagged “abort”	33

Table 2 shows the statistics of the corpus. 965 workers generated 270 dialogues which consists of 3,168 utterances. A dialogue was made in about 183 hours on average, however, we cannot discuss the detail about duration and interval time since they depend on workers’ availability at the period. While 33% dialogues (88/270) were judged as inadequate by workers, much fewer proportion of dialogues (4.5%) were judged as inadequate by annotators in our research laboratory as shown in Table 4. This seems to be caused by the easy work for a worker to check “inadequate” comparing to generating an utterance. This issue would be cleared up by additional instructions that gives workers the same amount of load such as having them to write the reason when they check “inadequate”.

Table 3 shows the length of dialogue for each 10 backgrounds. Each background has 27 dialogues whose length are diverse as shown in the table. The mean length has statistically significant difference (in T-test) between some backgrounds such as 1-8 and 2-3 while the SD has not significant difference (in F-test). It is intuitive that a background affects the length of dialogue. And the pseudo dialogue in which multiple workers made an utterance independently brought the intuitive result.

Table 3 Length (# of utterance) of dialogue in each background

Background ID	1	2	3	4	5	6	7	8	9	10
Mean of length	10.26	12.33	9.22	13.67	10.63	11.78	10.89	15.15	11.22	12.19
SD of length	5.95	5.48	5.09	5.20	5.49	5.81	5.19	4.37	5.90	6.04

The following is an instance of pseudo dialogue.

Agent: What kind of property are you looking for?

Customer: I prefer one which has a spacious kitchen and some cooking ranges because I want to tackle cooking.

Agent: How many burners do you want in the range?

Customer: Three, if possible.

Agent: Do you have any particular wishes for the range? For example, an IH cooking devise or a gas range.

Customer: I’d like saving and safe. Which do you recommend, gas range or IH devise?

Agent: If you are eager for cooking, I recommend a gas range, but IH devise is better from a viewpoint of safety. Do you have a child?

Customer: No. But The kitchen in my parent's house has IH devise and it's easy to use. So, I want a kitchen with two IH devises and one normal electric heat range if possible.

. . .

4 Comparison between Pseudo Dialogue and Chat Dialogue

Assigning more than one worker to a dialogue participant may cause some negative effect in the generated corpus. For the investigation we compare the pseudo dialogue corpus with the chat dialogue corpus. Table 4 shows the comparison between them.

The pseudo dialogue has less utterances in a dialogue and more characters in an utterance. In pseudo dialogue, workers cannot control following utterances. It prevent workers from plotting dialogue strategies such as planning utterances in the future turns. Hence workers would be forced to describe most of contents in which s/he could tell at that point. The constraint seems to cause the few numbers of long utterances. There is no significant difference between pseudo dialogue and chat dialogue with respect to occurrence of common vocabulary between two roles.

Table 4 The comparison between pseudo dialogue and chat dialogue

	Pseudo	Chat
average # of utterances in a dialogue	11.73	19.30
average # of characters in an utterance	27.57	14.99
average # of words in a dialogue	88.10	75.85
average # of common words in a dialogue	22.72	17.11
proportion of common words	0.246	0.223
average # of topics in a dialogue	4.644	7.011
average # of utterances in a topic	2.808	2.740
inadequate dialogue ratio (average in four annotators)	0.045	0.045

We had one annotator to annotate topic boundaries and then investigated the difference between the dialogues. The topics are features used often in property search such as “room layout” and “rent money”. The number of topic in a dialogue shows the same trend with “average # of utterance”. However, the number of utterances in a topic has no significant difference.

We also had four annotators to annotate inadequate dialogue for 50 sampled dialogues each. The result shows that the quality in terms of acceptance has no difference between them. However the agreement between annotators are not enough and we stooped the annotation. It is an issue to define “inadequacy” of dialogue.

5 Conclusion

We proposed a methodology to create a dialogue corpus assigning multiple crowd workers to one of two participants. It is difficult to apply tasks depending on other workers like dialogue generation to a crowdsourcing service. We addressed this issue by treating the task as a set of individual tasks in which a worker generate an utterance given a background and an utterances sequence history. Once a task is independent to be widely distributed, this method enables the corpus development to be scalable by letting us to avoid the labor of arranging and managing worker pairs.

The method brought some difference between pseudo dialogue and chat dialogue. For example, some phenomena, such as canceling previous utterance and successive utterances of same participant, were not seen in our dialogue data. However the analysis showed that the dialogue is not inadequate.

Our method is not applicable to every type of dialogue. It is necessary that workers can share a background, a goal and knowledge for the dialogue. In the setting of this study, a motivation of a customer is shared among workers and the required knowledge is easy to be confined. It is difficult to adapt our method to a chitchat which we cannot predict required knowledge of.

We constructed a small corpus and evaluated the method in this paper. For future works, we would construct a large corpus and evaluate the method quantitatively. For the reason of comparison, we also plan to develop a dialogue corpus in real time manner (namely the 1st method in Table 1) in which overlapping and cross-reference of topic may be found.

References

1. Higashinaka, R., Funakoshi, K., Araki, M., Tsukahara, H., Kobayashi, Y., Mizukami, M.: Towards taxonomy of errors in chat-oriented dialogue systems. In: Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 87–95 (2015)
2. Lasecki, W.S., Song, Y.C., Kautz, H., Bigham, J.P.: Real-time crowd labeling for deployable activity recognition. In: Proceedings of the 2013 conference on Computer supported cooperative work, pp. 1203–1212. ACM (2013)
3. Mitchell, M., Bohus, D., Kamar, E.: Crowdsourcing language generation templates for dialogue systems. In: Proceedings of the INLG and SIGDIAL 2014 Joint Session, pp. 16–24 (2014)
4. Novikova, J., Lemon, O., Rieser, V.: Crowd-sourcing nlg data: Pictures elicit better data. In: Proceedings of the 9th International Natural Language Generation conference, pp. 265–273 (2016)
5. Otani, N., Baba, Y., Kashima, H.: Quality control of crowdsourced classification using hierarchical class structures. Expert Systems with Applications **58**(1), 155–163 (2016)
6. Paperno, D., Kruszewsk, G., Lazaridou, A., Pham, Q.N., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., Fernández, R.: The lambada dataset: Word prediction requiring a broad discourse context. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1525–1534 (2016)
7. Sugiyama, H., Meguro, T., Higashinaka, R., Minami, Y.: Open-domain utterance generation for conversational dialogue systems using web-scale dependency structures. In: Proceedings of the SIGDIAL 2013 Conference, pp. 334–338 (2013)